The London-Lund Corpus 2

A new corpus of spoken British English in the making

Pöldvere, Nele; Paradis, Carita; Johansson, Victoria

2017

Document Version:
Publisher's PDF, also known as Version of record

The London-Lund Corpus 2:
A new corpus of spoken British English in the making

Nele Põldvere, Carita Paradis and Victoria Johansson
Lund University

This methodological paper describes and critically examines the major stages of building a spoken language corpus by reporting on the process of compiling the London-Lund Corpus 2 (LLC 2) of spoken British English. LLC 2 is of the same size and compiled on the same principles as its predecessor, the London-Lund Corpus (LLC 1), launched in 1975 (Svartvik, 1990). LLC 2 will allow linguists to study contemporary speech in different contexts and in combination with LLC 1 it will also allow us to study language change in a principled way using a comparable data set of the language spoken 50 years ago.

While LLC 2 shares important traits with LLC 1, efforts are made to also make it synchronically representative (see Leech, 2007 for the incompatible relation between comparability and representativeness). On the one hand, LLC 2 is similar to LLC 1 in that priority is given to spontaneous face-to-face conversation, the most basic type of language use (Clark, 1996), and data retrieved from public resources (e.g. broadcast interviews and parliamentary debates). On the other hand, LLC 2 differs from LLC 1 in that it includes computer-mediated communication, more specifically, Skype conversations, in order to represent speech situations that use technologies characteristic of the 21st century.

The compilation of LLC 2 entails the completion of four fundamental stages (modified from Thompson, 2004):

1. Data collection (recordings of spoken communication)
2. Transcription of the recordings
3. Markup and annotation
4. Access to the corpus

First, similar to LLC 1, the majority of the spontaneous face-to-face conversations are recorded at the University College London with native speakers of British English. Detailed information about the speakers is obtained in order to support sociolinguistic analyses of the data. At the transcribing stage, the recordings are turned into written form following a detailed transcription scheme. The scheme is largely based on the International Corpus of English; however, a number of modifications have been made. For instance, transcriptions in LLC 2 also include timestamps that connect each speaker turn to the corresponding location in the audio file, allowing for prosodic analyses of the conversations. Stage 3 entails the computerization of the transcriptions. Hardie's (2014) *Modest XML for Corpora* is followed in the encoding of the corpus for the purposes of distribution and archiving. Furthermore, in contrast to LLC 1, anonymisation is carried out not only in the transcriptions but also in the audio files themselves by altering the tonal patterns of names. This means that the audio files will be made available to the public alongside the timestamped transcriptions from the Lund University Humanities Lab's server (Stage 4).

Research on language variation and change is crucial for our understanding of the forces, motivations and mechanisms that languages are constantly subject to in communication. Without having access to large and modern computerized language resources, and especially to spoken language data in which change features prominently, these endeavours cannot be pursued. The compilation of LLC 2 will fill this gap and facilitate principled research in the field.

References

Clark, H. (1996). *Using language*. Cambridge: Cambridge University Press.

Hardie, A. (2014). Modest XML for corpora: Not a standard, but a suggestion. *ICAME Journal 38*, 73–103.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the Web* (pp. 133–149). Amsterdam: Rodopi.

Svartvik, J. (Ed). (1990). The London-Lund Corpus of spoken English: Description and research. Lund: Lund University Press.

Thompson, P. (2004). Spoken language corpora. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books.